CAUSAL RATE DISTORTION FUNCTION AND RELATIONS TO FILTERING THEORY

PHOTIOS A. STAVROU* AND CHARALAMBOS D. CHARALAMBOUS†

Abstract. A causal rate distortion function (RDF) is defined, existence of extremum solution is described via weak*-convergence, and its relation to filtering theory is discussed. The relation to filtering is obtained via a causal constraint imposed on the reconstruction kernel to be realizable while the extremum solution is given for the stationary case.

Key words. causal rate distortion function (RDF), realization, causal filter, weak*-convergence, optimal reconstruction kernel

AMS subject classifications. 28A33, 46E15, 60G07

1. Introduction. Shannon's information theory for reliable communication evolved over the years without much emphasis on real-time realizability or causality imposed on the communication sub-systems. In particular, the classical rate distortion function (RDF) for source data compression deals with the characterization of the optimal reconstruction conditional distribution subject to a fidelity criterion [1], without regard for realizability. Hence, coding schemes which achieve the RDF are not realizable.

On the other hand, filtering theory is developed by imposing real-time realizability on estimators with respect to measurement data. Although, both reliable communication and filtering (state estimation for control) are concerned with reconstruction of processes, the main underlying assumptions characterizing them are different.

In this paper, the intersection of rate distortion function (RDF) and realizable filtering theory is established by invoking the additional assumption that the reconstruction kernel is realizable via causal operations, while the optimal causal reconstruction kernel is derived. Consequently, the connection between causal RDF, its characterization via the optimal reconstruction kernel, and realizable filtering theory are established under very general conditions on the source (including Markov sources). The fundamental advantage of the new filtering approach based on causal RDF, is the ability to ensure average or probabilistic bounds on the estimation error, which is a non-trivial task when dealing with Bayesian filtering techniques.

The first relation between information theory and filtering via distortion rate function is discussed by R. S. Bucy in [2], by carrying out the computation of a realizable distortion rate function with square criteria for two samples of the Ornstein-Uhlenbeck process. The earlier work of A. K. Gorbunov and M. S. Pinsker [7] on ϵ-entropy defined via a causal constraint on the reproduction distribution of the RDF, although not directly related to the realizability question pursued by Bucy, computes the causal RDF for stationary Gaussian processes via power spectral densities. The realizability constraints imposed on the reproduction conditional distribution in [2] and [7] are different. The actual computation of the distortion rate or RDF in these works is based on the Gaussianity of the process, while no general theory is developed to handle arbitrary processes.

---

*Ph.D. student at ECE Department, University of Cyprus, Green Park, Aglantzias 91, P.O. Box 20537, 1687, Nicosia, Cyprus (`photios.stavrou@ucy.ac.cy`).

†Professor at ECE Department, University of Cyprus, Green Park, Aglantzias 91, P.O. Box 20537, 1687, Nicosia, Cyprus (`chadcha@ucy.ac.cy`).

The main results described are the following.

1) Existence of the causal RDF using the topology of weak*-convergence.
2) Closed form expression of the optimal reconstruction conditional distribution for stationary processes, which is realizable via causal operations.
3) Realization procedure of the filter based on the causal RDF.

Next, we give a high level discussion on Bayesian filtering theory and we present some aspects of the problem and results pursued in this paper. Consider a discrete-time process $X^n \triangleq \{X_0, X_1, \ldots, X_n\} \in \mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$, and its reconstruction $Y^n \triangleq \{Y_0, Y_1, \ldots, Y_n\} \in \mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$, where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are Polish spaces (complete separable metric spaces). The objective is to reconstruct X^n by Y^n causally subject to a distortion or fidelity criterion.

In classical filtering, one is given a mathematical model that generates the process X^n, $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \ldots, n\}$ often induced via discrete-time recursive dynamics, a mathematical model that generates observed data obtained from sensors, say, Z^n, $\{P_{Z_i|Z^{i-1},X^i}(dz_i|z^{i-1}, x^i) : i = 0, 1, \ldots, n\}$ while Y^n are the causal estimates of some function of the process X^n based on the observed data Z^n. Thus, in classical filtering theory both models which generate the unobserved and observed processes, X^n and Z^n, respectively, are given á priori. Fig. 1.1 illustrates the cascade block diagram of the filtering problem.
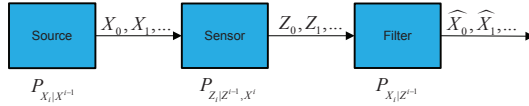

FIG. 1.1. *Block Diagram of Filtering Problem*

In causal rate distortion theory one is given the process X^n, which induces $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \ldots, n\}$, and determines the causal reconstruction conditional distribution $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \ldots, n\}$ which minimizes the mutual information between X^n and Y^n subject to a distortion or fidelity constraint, via a causal (realizability) constraint. The filter $\{Y_i : i = 0, 1, \ldots, n\}$ of $\{X_i : i = 0, 1, \ldots, n\}$ is found by realizing the reconstruction distribution $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \ldots, n\}$ via a cascade of sub-systems as shown in Fig. 1.2.
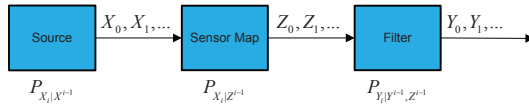

FIG. 1.2. *Block Diagram of Filtering via Causal Rate Distortion Function*

The distortion function or fidelity constraint between x^n and its reconstruction y^n, is a measurable function defined by

$$d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0, \infty], \ d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$$

The mutual information between X^n and Y^n, for a given distribution $P_{X^n}(dx^n)$, and conditional distribution $P_{Y^n|X^n}(dy^n|x^n)$, is defined by

$$(1.1) \quad I(X^n; Y^n) \triangleq \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{P_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n)$$

Define the $(n+1)-$fold causal convolution measure

$$(1.2) \qquad \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) - a.s.$$

The realizability constraint for a causal filter is defined by

$$(1.3) \quad \overrightarrow{Q}_{ad} \triangleq \left\{ P_{Y^n|X^n}(dy^n|x^n) : \ P_{Y^n|X^n}(dy^n|x^n) = \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) - a.s. \right\}$$

The realizability condition (1.3) is necessary, otherwise the connection between filtering and realizable rate distortion theory cannot be established. This is due to the fact that $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) - a.s.$, and hence in general, for each $i = 0, 1, \ldots, n$, the conditional distribution of Y_i depends on future symbols $\{X_{i+1}, X_{i+2}, \ldots, X_n\}$ in addition to the past and present symbols $\{Y^{i-1}, X^i\}$.
Causal Rate Distortion Function. The causal RDF is defined by

$$(1.4) \qquad R^c_{0,n}(D) \triangleq \inf_{P_{Y^n|X^n}(dy^n|x^n) \in \overrightarrow{Q}_{ad}: \ E\left\{d_{0,n}(X^n,Y^n) \leq D\right\}} I(X^n;Y^n)$$

Note that realizability condition (1.3) is different from the realizability condition in [2], which is defined under the assumption that Y_i is independent of $X^*_{j|i} \triangleq X_j - \mathbb{E}\left(X_j|X^i\right), j = i+1, i+2, \ldots,$. The claim here is that realizability condition (1.3) is more natural and applies to processes which are not necessarily Gaussian having square error distortion function. Realizability condition (1.3) is weaker than the causality condition in [7] defined by $X^\infty_{n+1} \leftrightarrow X^n \leftrightarrow Y^n$ forms a Markov chain.
The point to be made regarding (1.4) is that (see also Lemma 2.3):

$$P_{Y^n|X^n}(dy^n|x^n) = \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) - a.s. \Longleftrightarrow$$

$$I(X^n;Y^n) = \int \log\left(\frac{\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)}\right) \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) P_{X^n}(dx^n) \equiv \mathbb{I}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \quad (1.5)$$

where $\mathbb{I}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ points out the functional dependence of $I(X^n;Y^n)$ on $\{P_{X^n}, \overrightarrow{P}_{Y^n|X^n}\}$.
The paper is organized as follows. Section 2 discusses the formulation on abstract spaces. Section 3 establishes existence of optimal minimizing kernel, and Section 4 derives the stationary solution. Section 5 describes the realization of causal RDF. Throughout the manuscript proofs are omitted due to space limitation.

2. Problem Formulation. Let $\mathbb{N}^n \triangleq \{0, 1, \ldots, n\}$, $n \in \mathbb{N} \triangleq \{0, 1, 2, \ldots\}$. The source and reconstruction alphabets, respectively, are sequences of Polish spaces $\{\mathcal{X}_t : t \in \mathbb{N}\}$ and $\{\mathcal{Y}_t : t \in \mathbb{N}\}$, associated with their corresponding measurable spaces $(\mathcal{X}_t, \mathcal{B}(\mathcal{X}_t))$ and $(\mathcal{Y}_t, \mathcal{B}(\mathcal{Y}_t))$, $t \in \mathbb{N}$. Sequences of alphabets are identified with the product spaces $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n})) \triangleq \times_{k=0}^{n}(\mathcal{X}_k, \mathcal{B}(\mathcal{X}_k))$, and $(\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n})) \triangleq \times_{k=0}^{n}(\mathcal{Y}_k, \mathcal{B}(\mathcal{Y}_k))$. The source and reconstruction are processes denoted by $X^n \triangleq \{X_t : t \in \mathbb{N}^n\}$, $X : \mathbb{N}^n \times \Omega \mapsto \mathcal{X}_t$, and by $Y^n \triangleq \{Y_t : t \in \mathbb{N}^n\}$, $Y : \mathbb{N}^n \times \Omega \mapsto \mathcal{Y}_t$, respectively. Probability measures on any measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ are denoted by $\mathcal{M}_1(\mathcal{Z})$. It is assumed that the σ-algebras $\sigma\{X^{-1}\} = \sigma\{Y^{-1}\} = \{\emptyset, \Omega\}$.
DEFINITION 2.1. *Let $(\mathcal{X}, \mathcal{B}(\mathcal{X})), (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ be measurable spaces in which $\mathcal{Y}$ is a Polish Space. A stochastic kernel on $\mathcal{Y}$ given $\mathcal{X}$ is a mapping $q : \mathcal{B}(\mathcal{Y}) \times \mathcal{X} \to [0,1]$ satisfying the following two properties:*

1) *For every $x \in \mathcal{X}$, the set function $q(\cdot; x)$ is a probability measure (possibly finitely additive) on $\mathcal{B}(\mathcal{Y})$.*

2) *For every $F \in \mathcal{B}(\mathcal{Y})$, the function $q(F; \cdot)$ is $\mathcal{B}(\mathcal{X})$-measurable.*

The set of all such stochastic Kernels is denoted by $\mathcal{Q}(\mathcal{Y}; \mathcal{X})$.

DEFINITION 2.2. *Given measurable spaces $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n}))$, $(\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n}))$, then*

1) *A Non-Causal Data Compression Channel is a stochastic kernel $q_{0,n}(dy^n; x^n) \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ which admits a factorization into a non-causal sequence*

$$q_{0,n}(dy^n; x^n) = \otimes_{i=0}^n q_i(dy_i; y^{i-1}, x^n)$$

where $q_i(dy_i; y^{i-1}, x^n) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,n}), i = 0, \ldots, n, \ n \in \mathbb{N}$.

2) *A Causally Restricted Data Compression Channel is a stochastic kernel $q_{0,n}(dy^n; x^n) \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ which admits a factorization into a causal sequence*

$$q_{0,n}(dy^n; x^n) = \otimes_{i=0}^n q_i(dy_i; y^{i-1}, x^i) - a.s.,$$

where $q_i \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i}), i = 0, \ldots, n, \ n \in \mathbb{N}$.

2.1. Causal Rate Distortion Function. In this subsection the causal RDF is defined. Given a source probability measure $\mu_{0,n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$ (possibly finite additive) and a reconstruction Kernel $q_{0,n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$, one can define three probability measures as follows.

(P1): The joint measure $P_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n} \times \mathcal{X}_{0,n})$:

$$P_{0,n}(G_{0,n}) \triangleq (\mu_{0,n} \otimes q_{0,n})(G_{0,n}), \ G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n})$$
$$= \int_{\mathcal{X}_{0,n}} q_{0,n}(G_{0,n,x^n}; x^n)\mu_{0,n}(dx^n)$$

where G_{0,n,x^n} is the x^n-section of $G_{0,n}$ at point x^n defined by $G_{0,n,x^n} \triangleq \{y^n \in \mathcal{Y}_{0,n} : (x^n, y^n) \in G_{0,n}\}$ and $\otimes$ denotes the convolution.

(P2): The marginal measure $\nu_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$:

$$\nu_{0,n}(F_{0,n}) \triangleq P_{0,n}(\mathcal{X}_{0,n} \times F_{0,n}), \ F_{0,n} \in \mathcal{B}(\mathcal{Y}_{0,n})$$
$$= \int_{\mathcal{X}_{0,n}} q_{0,n}((\mathcal{X}_{0,n} \times F_{0,n})_{x^n}; x^n)\mu_{0,n}(dx^n) = \int_{\mathcal{X}_{0,n}} q_{0,n}(F_{0,n}; x^n)\mu_{0,n}(dx^n)$$

(P3): The product measure $\pi_{0,n} : \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n}) \mapsto [0, 1]$ of $\mu_{0,n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$ and $\nu_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ for $G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n})$:

$$\pi_{0,n}(G_{0,n}) \triangleq (\mu_{0,n} \times \nu_{0,n})(G_{0,n}) = \int_{\mathcal{X}_{0,n}} \nu_{0,n}(G_{0,n,x^n})\mu_{0,n}(dx^n)$$

The precise definition of mutual information between two sequences of Random Variables X^n and Y^n, denoted $I(X^n; Y^n)$ is defined via the Kullback-Leibler distance (or relative entropy) between the joint probability distribution of (X^n, Y^n) and the product of its marginal probability distributions of X^n and Y^n, using the Radon-Nikodym derivative. Hence, by the chain rule of relative entropy:

$$I(X^n; Y^n) \triangleq \mathbb{D}(P_{0,n} || \pi_{0,n}) = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log\left(\frac{d(\mu_{0,n} \otimes q_{0,n})}{d(\mu_{0,n} \times \nu_{0,n})}\right) d(\mu_{0,n} \otimes q_{0,n})$$
$$= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log\left(\frac{q_{0,n}(dy^n; x^n)}{\nu_{0,n}(dy^n)}\right) q_{0,n}(dy^n; dx^n) \otimes \mu_{0,n}(dx^n)$$
$$(2.1) \qquad = \int_{\mathcal{X}_{0,n}} \mathbb{D}(q_{0,n}(\cdot; x^n) || \nu_{0,n}(\cdot))\mu_{0,n}(dx^n) \equiv \mathbb{I}(\mu_{0,n}, q_{0,n})$$

The next lemma relates causal product reconstruction kernels and conditional independence.

LEMMA 2.3. *The following are equivalent for each $n \in \mathbb{N}$.*

1) *$q_{0,n}(dy^n; x^n) = \overrightarrow{q}_{0,n}(dy^n; x^n)$-a.s., defined in Definition 2.2-2).*
2) *For each $i = 0, 1, \ldots, n-1$, $Y_i \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow (X_{i+1}, X_{i+2}, \ldots, X_n)$, forms a Markov chain.*
3) *For each $i = 0, 1, \ldots, n-1$, $Y^i \leftrightarrow X^i \leftrightarrow X_{i+1}$ forms a Markov chain.*

According to Lemma 2.3, for causally restricted kernels

$$I(X^n; Y^n) = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log\Big(\frac{\overrightarrow{q}_{0,n}(dy^n; x^n)}{\nu_{0,n}(dy^n)} \Big) \overrightarrow{q}_{0,n}(dy^n; dx^n) \otimes \mu_{0,n}(dx^n)$$

(2.2) $$\equiv \mathbb{I}(\mu_{0,n}, \overrightarrow{q}_{0,n})$$

where (2.2) states that $I(X^n; Y^n)$ is a functional of $\{\mu_{0,n}, \overrightarrow{q}_{0,n}\}$. Hence, causal RDF is defined by optimizing $\mathbb{I}(\mu_{0,n}, q_{0,n})$ over $q_{0,n} \in Q_{0,n}(D)$ where $Q_{0,n}(D) = \{q_{0,n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \int_{\mathcal{X}_{0,n}} \int_{\mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n) q_{0,n}(dy^n; x^n) \otimes \mu_{0,n}(dx^n) \leq D\}$ subject to the realizability constraint $q_{0,n}(dy^n; x^n) = \overrightarrow{q}_{0,n}(dy^n; x^n) - a.s.$, which satisfies a distortion constraint, or via (2.2).

DEFINITION 2.4. (*Causal Rate Distortion Function*) *Suppose $d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$, where $\rho_{0,i} : \mathcal{X}_{0,i} \times \mathcal{Y}_{0,i} \to [0, \infty)$, is a sequence of $\mathcal{B}(\mathcal{X}_{0,i}) \times \mathcal{B}(\mathcal{Y}_{0,i})$-measurable distortion functions, and let $\overrightarrow{Q}_{0,n}(D)$ (assuming is non-empty) denotes the average distortion or fidelity constraint defined by*

$$\overrightarrow{Q}_{0,n}(D) \triangleq Q_{0,n}(D) \bigcap \overrightarrow{Q}_{ad}, \ D \geq 0$$

The causal RDF associated with the causally restricted kernel is defined by

(2.3) $$R_{0,n}^c(D) \triangleq \inf_{q_{0,n} \in \overrightarrow{Q}_{0,n}(D)} \mathbb{I}(\mu_{0,n}, q_{0,n})$$

3. Existence of Optimal Causal Reconstruction Kernel. In this section, appropriate topologies and function spaces are introduced and existence of the minimizing causal product kernel in (2.3) is shown.

3.1. Abstract Spaces. Let $BC(\mathcal{Y}_{0,n})$ denote the vector space of bounded continuous real valued functions defined on the Polish space $\mathcal{Y}_{0,n}$. Furnished with the sup norm topology, this is a Banach space. The topological dual of $BC(\mathcal{Y}_{0,n})$ denoted by $\big(BC(\mathcal{Y}_{0,n})\big)^*$ is isometrically isomorphic to the Banach space of finitely additive regular bounded signed measures on $\mathcal{Y}_{0,n}$ [5], denoted by $M_{rba}(\mathcal{Y}_{0,n})$. Let $\Pi_{rba}(\mathcal{Y}_{0,n}) \subset M_{rba}(\mathcal{Y}_{0,n})$ denote the set of regular bounded finitely additive probability measures on $\mathcal{Y}_{0,n}$. Clearly if $\mathcal{Y}_{0,n}$ is compact, then $\big(BC(\mathcal{Y}_{0,n})\big)^*$ will be isometrically isomorphic to the space of countably additive signed measures, as in [4]. Denote by $L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ the space of all $\mu_{0,n}$-integrable functions defined on $\mathcal{X}_{0,n}$ with values in $BC(\mathcal{Y}_{0,n})$, so that for each $\phi \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ its norm is defined by

$$\| \phi \|_{\mu_{0,n}} \triangleq \int_{\mathcal{X}_{0,n}} ||\phi(x^n, \cdot)||_{BC(\mathcal{Y}_{0,n})} \mu_{0,n}(dx^n) < \infty$$

The norm topology $\| \phi \|_{\mu_{0,n}}$, makes $L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ a Banach space, and it follows from the theory of "lifting" [8] that the dual of this space is $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$, denoting the space of all $M_{rba}(\mathcal{Y}_{0,n})$ valued functions $\{q\}$ which are weak*-measurable in the sense that for each $\phi \in BC(\mathcal{Y}_{0,n})$, $x^n \to q_{x^n}(\phi) \triangleq \int_{\mathcal{Y}_{0,n}} \phi(y^n)q(dy^n; x^n)$ is $\mu_{0,n}$-measurable and $\mu_{0,n}$-essentially bounded.

3.2. Weak*-Compactness and Existence. Define an admissible set of stochastic kernels associated with classical RDF by

$$Q_{ad} \triangleq L_\infty^w(\mu_{0,n}, \Pi_{rba}(\mathcal{Y}_{0,n})) \subset L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$$

Clearly, Q_{ad} is a unit sphere in $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$. For each $\phi \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ we can define a linear functional on $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$ by

$$\ell_\phi(q_{0,n}) \triangleq \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} \phi(x^n, y^n)q_{0,n}(dy^n; x^n) \right) \mu_{0,n}(dx^n)$$

This is a bounded, linear and weak*-continuous functional on $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$.

For $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0,\infty)$ measurable and $d_{0,n} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ the distortion constraint set of the classical RDF is $Q_{0,n}(D) \triangleq \{q \in Q_{ad} : \ell_{d_{0,n}}(q_{0,n}) \leq D\}$.

LEMMA 3.1. *For $\ell_{d_{0,n}} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$, the set $Q_{0,n}(D)$ is weak*-bounded and weak*-closed subset of Q_{ad}.*

Hence $Q_{0,n}(D)$ is weak*-compact (compactness of Q_{ad} follows from Alaoglu's Theorem [5]).

LEMMA 3.2. *Let $\mathcal{X}_{0,n}, \mathcal{Y}_{0,n}$ be two Polish spaces and $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0,\infty]$, a measurable, nonnegative, extended real valued function, such that for a fixed $x^n \in \mathcal{X}_{0,n}$, $y^n \to d(x^n, \cdot)$ is continuous on $\mathcal{Y}_{0,n}$, for $\mu_{0,n}$-almost all $x^n \in \mathcal{X}_{0,n}$, and $d_{0,n} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$. For any $D \in [0,\infty)$, introduce the set*

$$Q_{0,n}(D) \triangleq \{q_{0,n} \in Q_{ad} : \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n)q_{0,n}(dy^n; x^n) \right) \mu_{0,n}(dx^n) \leq D\}$$

and suppose it is nonempty.
Then $Q_{0,n}(D)$ is a weak-closed subset of Q_{ad} and hence weak*-compact.*

Next, we define the realizability constraint via causally restricted kernels as follows

$$\overrightarrow{Q}_{ad} = \left\{ q_{0,n} \in Q_{ad} : q_{0,n}(dy^n; x^n) = \overrightarrow{q}_{0,n}(dy^n; x^n) - a.s. \right\}$$

which satisfy an average distortion function as follows:

$$\overrightarrow{Q}_{0,n}(D) \triangleq Q_{0,n}(D) \bigcap \overrightarrow{Q}_{ad}$$
$$= \left\{ q_{0,n} \in \overrightarrow{Q}_{ad} : \ell_{d_{0,n}}(\overrightarrow{q}_{0,n}) \triangleq \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n)\overrightarrow{q}_{0,n}(dy^n; x^n) \right) \right.$$
$$\left. \otimes \mu_{0,n}(dx^n) \right\}$$

The following is assumed.

ASSUMPTION 3.3. *Let $\mathcal{X}_{0,n}$ and $\mathcal{Y}_{0,n}$ be Polish spaces and $\overrightarrow{Q}_{ad}$ weak*-closed.*

REMARK 3.4. *The conditions 1) $\mathcal{Y}_{0,n}$ is a compact Polish space, and 2) for all $h(\cdot) \in BC(\mathcal{Y}_n)$, the function $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{Y}_n} h(y)q_n(dy; y^{n-1}, x^n) \in$*

$\mathbb{R}$ *is continuous jointly in the variables* $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1}$ *are sufficient for* $\overrightarrow{Q}_{ad}$ *to be weak*-closed.*

THEOREM 3.5. *Suppose Assumption 3.3 and the conditions of Lemma 3.2 hold. For any* $D \in [0, \infty)$, *introduce the set*

$$Q_{0,n}(D) \triangleq \{q_{0,n} \in \overrightarrow{Q}_{ad} : \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} d(x^n, y^n) \overrightarrow{q}_{0,n}(dy^n; x^n) \right) \mu_{0,n}(dx^n) \le D\}$$

and suppose it is nonempty.
Then $\overrightarrow{Q}_{0,n}(D)$ *is a weak*-closed subset of* $\overrightarrow{Q}_{ad}$ *and hence weak*-compact.*

THEOREM 3.6. *Under Theorem 3.5,* $R^c_{0,n}(D)$ *has a minimum.*

Proof. Follows from weak*-compactness of $\overrightarrow{Q}_{ad}$ and lower semicontinuity of $\mathbb{I}(\mu_{0,n}, q_{0,n})$ with respect to $q_{0,n}$ for a fixed $\mu_{0,n}$. □

4. Necessary Conditions of Optimality of Causal Rate Distortion Function. In this section the form of the optimal causal product reconstruction kernels is derived under a stationarity assumption. The method is based on calculus of variations on the space of measures [9].

ASSUMPTION 4.1. *The family of measures* $\overrightarrow{q}_{0,n}(dy^n; x^n) = \otimes_{i=0}^n q_i(dy_i; y^{i-1}, x^i) -$ *a.s., is the convolution of stationary conditional distributions.*

Assumption 4.1 holds for stationary process $\{(X_i, Y_i) : i \in \mathbb{N}\}$ and $\rho_{0,i}(x^i, y^i) \equiv \rho(T^i x^n, T^i y^n)$, where $T^i x^n$ is the shift operator on x^n. Utilizing Assumption 4.1, which holds for stationary processes and a single letter distortion function, the Gateaux differential of $\mathbb{I}(\mu_{0,n}, q_{0,n})$ is done in only one direction (since $q_i(dy_i; y^{i-1}, x^i)$ are stationary).

The constrained problem defined by (2.3) can be reformulated using Lagrange multipliers as follows (equivalence of constrained and unconstrained problems follows similarly as in [9]).

$$(4.1) \qquad R^c_{0,n}(D) = \inf_{q_{0,n} \in \overrightarrow{Q}_{ad}} \left\{ \mathbb{I}(\mu_{0,n}, q_{0,n}) - s(\ell_{d_{0,n}}(q_{0,n}) - D) \right\}$$

and $s \in (-\infty, 0]$ is the Lagrange multiplier.
Note that $\overrightarrow{Q}_{ad}$ is a proper subset of the vector space $L^w_\infty(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$ which represent the realizability constraint. Therefore, one should introduce another set of Lagrange multipliers to obtain an optimization on the vector space $L^w_\infty(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$ without constraints.

THEOREM 4.2. *Suppose* $d_{0,n}(x^n, y^n) = \sum_{i=0}^n \rho(T^i x^n, T^i y^n)$ *and the Assumption 3.3 holds. The infimum in (4.1) is attained at* $q^*_{0,n} \in L^w_\infty(\mu_{0,n}, \Pi_{rba}(\mathcal{Y}_{0,n})) \cap \overrightarrow{Q}_{ad}$ *given by*

$$(4.2) \qquad \begin{aligned} q^*_{0,n}(dy^n; x^n) &= \overrightarrow{q}^*_{0,n}(dy^n; x^n) - a.s. \\ &= \otimes_{i=0}^n q^*_i(dy_i; y^{i-1}, x^i) - a.s \\ &= \otimes_{i=0}^n \frac{e^{s\rho(T^i x^n, T^i y^n)} \nu^*_i(dy_i; y^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} \nu^*_i(dy_i; y^{i-1})}, \quad s \le 0 \end{aligned}$$

and $\nu^*_i(dy_i; y^{i-1}) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1})$. *The causal RDF is given by*

$$(4.3) \qquad \begin{aligned} R^c_{0,n}(D) &= sD - \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i-1}} \log \left(\int_{\mathcal{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} \nu^*_i(dy_i; y^{i-1}) \right) \\ &\times \overrightarrow{q}^*_{0,i-1}(dy^{i-1}; x^{i-1}) \otimes \mu_{0,i}(dx^i) \end{aligned}$$

If $R_{0,n}^c(D) > 0$ then $s < 0$ and

$$\sum_{i=0}^{n} \int_{\mathcal{X}_{0,i}} \int_{\mathcal{Y}_{0,i}} \rho(T^i x^n, T^i y^n) \overrightarrow{q}_{0,i}^*(dy^i; x^i) \mu_{0,i}(dx^i) = D$$

REMARK 4.3. *Note that if the distortion function satisfies $\rho(T^i x^n, T^i y^n) = \rho(x_i, T^i y^n)$ then $q_i^*(dy_i; y^{i-1}, x^i) = q_i^*(dy_i; y^{i-1}, x_i) - a.s., \ i \in \mathbb{N}^n$, that is, the reconstruction kernel is Markov in X^n.*

5. Realization of Causal Rate Distortion Function. Fig. 5.1 illustrates a cascade of sub-systems which realizes the causal RDF. This is called source-channel matching in information theory [6]. It is also described in [3] and [11] and is essential in control applications since this technique allows us to design encoding/decoding schemes without delays.
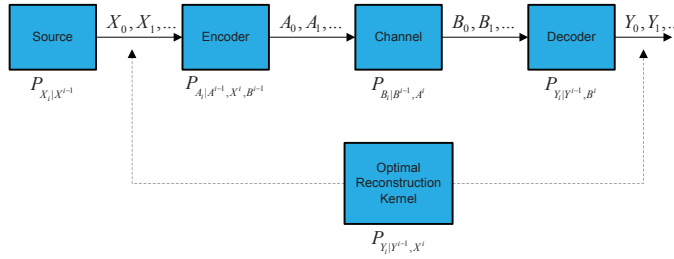

FIG. 5.1. *Block Diagram of Realizable Causal Rate Distortion Function*

Examples to illustrate the concepts can be found in [3, 10].

REFERENCES

[1] T. BERGER, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.
[2] R. S. BUCY, *Distortion rate theory and filtering*, IEEE Transactions on Information Theory, 28 (1982), pp. 336–340.
[3] C. D. CHARALAMBOUS AND A. FARHADI, *LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels*, Automatica, 44 (2008), pp. 3181–3188.
[4] I. CSISZÁR, *On an extremum problem of information theory*, Studia Scientiarum Mathematicarum Hungarica, 9 (1974), pp. 57–71.
[5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I: General Theory*, John Wiley & Sons, Inc., Hoboken, New Jersey, 1988.
[6] M. GASTPAR, B. RIMOLDI, AND M. VETTERLI, *To code, or not to code: Lossy source-channel communication revisited*, IEEE Transactions on Information Theory, 49 (2003), pp. 1147–1158.
[7] A. K. GORBUNOV AND M. S. PINSKER, *Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes*, Problems of Information Transmission, 27 (1991), pp. 361–365.
[8] A. IONESCU TULCEA AND C. IONESCU TULCEA, *Topics in the Theory of Lifting*, Springer-Verlag, Berlin, Heidelberg, New York, 1969.
[9] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley & Sons, Inc., New York, 1969.
[10] P. A. STAVROU, C. D. CHARALAMBOUS, AND C. K. KOURTELLARIS, *Realizable rate distortion function and Bayesian filtering theory*, submitted to IEEE Information Theory Workshop (ITW), abs/1204.2980 (2012).
[11] S. C. TATIKONDA, *Control Over Communication Constraints*, PhD thesis, Mass. Inst. of Tech. (M.I.T.), Cambridge, MA, 2000.